Australian Bureau of Statistics

1351.0.55.034

Research Paper

Sampling-Based Clerical Review Methods in Probabilistic Linking

www.abs.gov.au

New Issue

Research Paper

Sampling-Based Clerical Review Methods in Probabilistic Linking

Tenniel Guiver

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 25 MAY 2011

ABS Catalogue no. 1351.0.55.034

© Commonwealth of Australia 2011

This work is copyright. Apart from any use as permitted under the *Copyright Act 1968*, no part may be reproduced by any process without prior written permission from the Commonwealth. Requests and inquiries concerning reproduction and rights in this publication should be addressed to The Manager, Intermediary Management, Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616, by telephone (02) 6252 6998, fax (02) 6252 7102, or email <intermediary.management@abs.gov.au>.

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Dr Shiji Zhao, Analytical Services Branch on Canberra (02) 6252 6832 or email <analytical.services@abs.gov.au>.

CONTENTS

SAMPLING-BASED CLERICAL REVIEW METHODS IN PROBABILISTIC LINKING

Tenniel Guiver
Analytical Services

ABSTRACT

Probabilistic data linking aims to link records that are believed to belong to the same person from two different data sets. Candidate record pairs are given a weight based on the degree of agreement between fields on the two records. Record pairs with a weight above some upper cut-off are declared links while those with a weight below some lower cut-off are declared non-links. However, there are many record pairs that cannot be automatically assigned a status and are designated for clerical review. Clerical review is a time-consuming and resource-intensive stage of the data linking process requiring careful visual inspection and keyboard use. Acceptance sampling is proposed to dramatically reduce the amount of clerical inspection. Sampling is also proposed as a method to provide an accurate and reliable means of assessing and setting the most appropriate clerical review bounds.

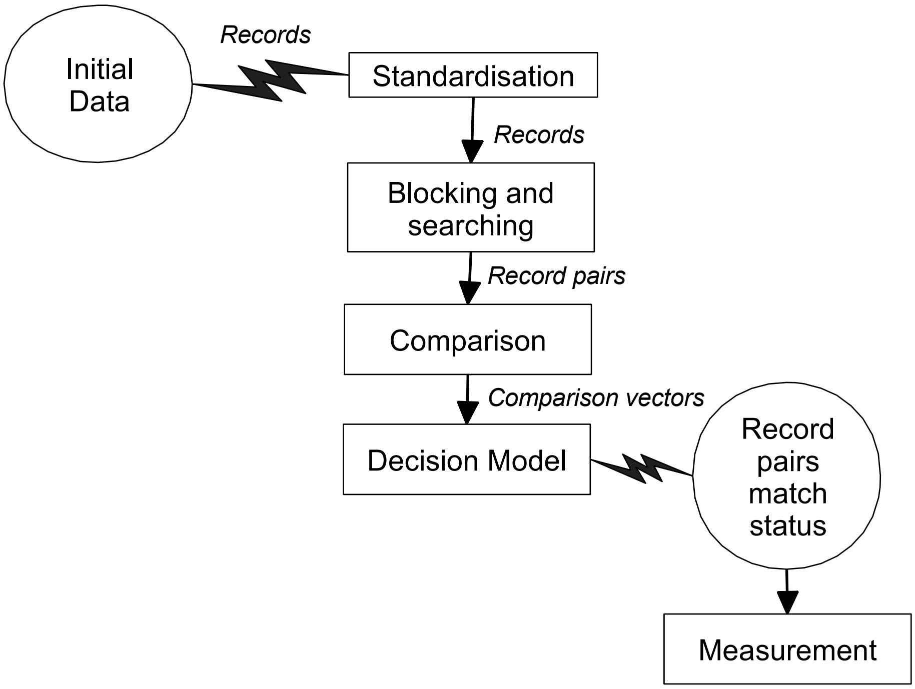
1. BACKGROUND

Fellegi and Sunter (1969) proposed a probabilistic framework to link records together using several fields. Conn and Bishop (2005) explored this framework, conducted data linking trials and proposed a number of studies using this methodology. This framework has been subsequently applied in a number of linking studies by the ABS as part of the Census Data Enhancement project. Further information about the range of studies in the Census Data Enhancement project is available in ABS publications (ABS, 2005a, 2005b, 2006, 2009).

A key feature of the probabilistic data linking framework is the ability to handle a variety of linking variables and record comparison methods to come up with a single numerical measure of how well two particular records link. [1] This allows ranking of all possible links and optimal assignment of link status. The assignment of link or non-link status is generally performed using a combination of automatic assignment and manual inspection or clerical review.

[1] In this paper the term 'link' is used to describe two records that have been linked by the data linking process whereas 'match' is used to mean that two records belong to the same entity,

1.1 The generalised data linking process



This framework can be generalised into the following steps (illustrated in figure 1.1):

- standardisation,
- blocking and searching,
- record pair comparisons and
- a decision model.

Standardisation removes inconsistencies and parses text fields so that the data items in each file are comparable.

Blocking reduces the number of comparisons needed by only comparing record pairs where links are more likely to be found. Records on each file are placed into blocks so that only record pairs that agree on certain data items are compared.

During the comparison stage, each linkage field for a record pair is compared and the level of agreement is measured. The probabilities of agreement of fields conditional upon whether the record pair is a match or non-match along with a wide range of field comparators are used to determine a field comparison weight. These field comparison weights are summed over the linking variables to form a record pair comparison weight.

Finally, a decision rule based on cut-offs determines whether the record pair is linked, not linked or considered further as a possible link. The decision rule is based on the record pair comparison weight and is used to assign a link status. The simplest decision rule is a single cut-off weight, all record pairs with a weight equal to or above this weight are assigned as links and all those below are assigned as non-links. A more sophisticated decision rule has lower and upper cut-off. Record pairs with a weight above the upper cut-off are declared links while those with a weight below some lower cut-off are declared non-links. The record pairs that cannot be automatically assigned a status are designated for clerical review. Clerical review involves human assessment of each record pair to resolve link status.

2. TRADITIONAL METHODOLOGIES FOR CLERICAL REVIEW

Record pairs considered for clerical review do not have sufficient evidence to assign them as either a link or non-link and human assessment of these record pairs are required. There are a number of challenges involved with human assessment:

- the setting of cut-offs for clerical review;
- the number of record pairs requiring human assessment; and
- the availability of data fields suitable for human assessment.

2.1 Setting cut-offs for clerical review

Before clerical review can take place, the upper and lower cut-offs need to be determined. These cut-offs define the record pair comparison weight range of the record pairs to be clerically reviewed.

The entire set of record pair comparisons can be examined by eye. In general, there will be far too many record pairs to be fully examined; so a sample of record pairs may be used. Spot checking the record pairs involves selecting an ad hoc sample over a range of comparison weights and assessing the likelihood of correct linkage. It is not possible to make any firm quantitative conclusions about the quality of the record pairs using this type of sampling. Sampling using a probability based design allows for an informed decision about quality and has been suggested by other researchers. Gill (2001) suggests taking a sample and using observations on this sample to select cut-offs. The Centre for Health Record Linkage, CHeRel (2008), suggests setting cut-offs and then selecting a sample of linked record pairs and a sample of unlinked record pairs and adjusting the cut-offs up or down on the basis of a set of rules using the observed false positive and true positive rates.

2.2 Clerically reviewing record-pairs

Once the upper and lower cut-offs have been set the record pairs can be reviewed. Each identified record pair requires manual inspection. If the reviewer finds evidence the pair is more likely to be a match then the record pair is assigned as a link, otherwise the record pair is assigned as a non-link. Jaro (1989) and Gill (2001) describe clerical review using manual inspection of computer generated forms. These forms contain the data fields from each file as well as record identifiers. More commonly, clerical review is performed using a computer based user interface. Christen (2008) describes a prospective module for the FEBRL record linkage system that displays the fields from both files side by side in a graphical user interface, allowing an operator to assess whether or not the record pair is more likely to be a match or non-match. The benefit of a user interface integrated in the linking software is that the clerically assigned link statuses can be directly fed back into the linking process.

The assessment of record pairs is a subjective process. Additionally, each linkage project generally presents new issues relating to data items and data quality. For example, there are feasible and infeasible changes in data items; marital status can feasibly change from never married to married, but the reverse is not a feasible change (though may represent an error in data reporting or processing). Tolerances for numeric variables can be established; for example a one or two year difference in age may be considered to be acceptable. If text data has been processed by Optical Character Recognition (OCR) then typical errors (for example Ms being misinterpreted as Hs) can be anticipated.

The volume of clerical review pairs can generate a significant workload for staff and should not be underestimated. In one study during the Census Data Enhancement project the number of record pairs identified for clerical review was 131,000.

2.3 Data fields used for clerical review

Name and address are the most useful fields during clerical review. These fields can be subject to typographic errors, reporting errors or data processing errors. Porter and Winkler (1997) discuss a range of comparators for assessing the similarity of text strings in the probabilistic data linking process. However, there are many types of error that cannot be automatically detected and accounted for by these *approximate string comparators*. Two character strings each representing a name may have low level agreement when compared using an approximate string comparator. However a person performing clerical review might conclude that these fields do provide evidence that the records belong to the same person.

2.1 Name fields potentially belonging to the same person that could be identified by human review

Name recorded on file 1	*Name recorded on file 2*	*Reason records may belong to same person*
JENNIFER	IFHHIEER	OCR has replaced characters with other characters of a similar shape
XXXXXTIM	TIM	Extraneous characters have been inserted in front of an otherwise valid and consistent name
ELIZABETH	BESS	Bess is a common diminutive form of Elizabeth

The examples in table 2.1 would have relatively low comparison scores when using an approximate string comparator and may contribute to evidence of the records being classified as a non-link. However, a human reviewer could use these fields as evidence of the records being a match.

Studies in the Census Data Enhancement Project, as well as other studies such as AIHW (2003) and Karmel (2004), have conducted data linkage without using name and address. A challenge of undertaking data linking without name and address is the lack of fields with which to conduct meaningful clerical review. Examples of other fields that may typically be available for a linkage project are listed in table 2.2.

2.2 Example fields available in a linking study not using name and address

Date of birth
Sex
Indigenous status
Country of birth
Ancestry
Geographic code (e.g. Census Collection District or Mesh Block)
Language spoken
Year of arrival
Marital status
Religion
Level of highest qualification
Field of study
Occupation
Industry

Most of these fields are either numeric or can be coded to numeric values, and have some use in clerical review. For example, country of birth from one file could be compared to ancestry or language on the second file. However, such comparisons require significantly more effort and consideration to determine a link status. The reviewer will have to consider the weight of evidence that agreement and disagreement on a range of fields provide. The reviewer will also have to assess feasibility of changes in data items and may have to consider the compatibility between different occupations and fields of study.

Several quality studies in the Census Data Enhancement project involved linking without name and address. Because of the complexity of clerical review when these fields are not present, those linkages used a single cut-off. Record pairs above this cut-off were assigned as a link, while those below the cut-off were assigned as non-links.

3. METHODOLOGIES FOR SAMPLING-BASED CLERICAL REVIEW

3.1 Uses of sampling in clerical review

In the context of clerical review in data linking, sampling can be applied by forming batches of record pairs and taking a sample of record pairs from each batch. Each sample of record pairs can then be clerically reviewed, rather than reviewing the entire batch.

These results can be used in a number of ways. Acceptance sampling can be used to assign link status to batches automatically. Alternatively the results of the sampling can be used to establish a profile of the estimated links and non-links by record pair comparison weight. This profile can then be used to establish cut-offs.

3.2 Overview of acceptance sampling

Acceptance sampling is a process that replaces exhaustive inspection of accepting or rejecting the quality of items. The process has many applications in industry and other fields. For example:

- In manufacturing, each item may be a ball bearing. A ball bearing may be accepted or rejected on the basis of physical dimensions.
- In administration, an item may be a completed form. An administrative form can be accepted or rejected on the basis of completion compliance.
- In clerical review, the item is a record pair.

Additional example applications are given by Montgomery (2005) and Juran and Godfrey (1999). In the context of data linkage, a record pair is assessed and consequently sentenced on the basis of whether it can be assigned as a link or a non-link.

Acceptance sampling is a well established methodology and is the subject of a number of Australian and International standards. Acceptance sampling uses sampling to reduce the amount of inspection. The items produced are divided into batches (or lots). A sample is selected from each batch and each item selected is inspected. The number of rejectable items in the batch is compared to a critical value. The entire batch is sentenced on the basis of observations made on the sample. If the number of rejectable items observed is greater than or equal to the critical value then the batch is rejected, otherwise the batch is accepted.

To define a sampling scheme the analyst must specify:

- The acceptable quality level (AQL). This is the quality level above which we would reject a batch with only a small defined level of probability.

- The producer's risk. This is the probability of rejecting an 'acceptable' batch, and is generally set low (say, 5–10%).
- The rejectable quality level (RQL). This is the quality level below which we wish to reject with a defined high level of probability.
- The consumer's risk. This is the probability of not rejecting a 'rejectable' batch, and is generally set low, so that the probability of rejecting a rejectable batch (or power) is high (say, 80–90%).

These parameters are used to derive a sampling scheme. The simplest sampling scheme, a single sampling scheme, comprises a sample size and a critical value.

3.3 Automatically sentencing batches

In classical acceptance sampling each batch is sentenced on the basis of the sample. In the context of clerical review a batch can be:

- accepted as a batch of links containing a tolerable number of non-matches;
- accepted as a batch of non-links containing a tolerable number of matches; or
- sent to clerical review.

Effectively, each batch of record pairs is simultaneously subjected to two tests. One test determines whether the batch can be accepted as links and a second test determines whether the batch can be accepted as non-links. To reflect this, a batch will be said to be 'assigned as links', 'assigned as non-links' or 'sent to clerical review' rather than be 'accepted' or 'rejected'.

The steps involved in acceptance sampling based clerical review are as follows:

1. Examine the frequency distribution of record pair comparison weights. It is not necessary to sample the entire range as record pairs with very low comparison weights are very unlikely to be matches while record pairs with very high comparison weights are very likely to be matches.
2. Determine an acceptance sampling scheme (refer to section 3.5).
3. The record pairs are ordered by record pair comparison weight and divided into batches. This can be done on the basis of equal weight ranges or equal batch sizes. A greater number of batches will result in a greater amount of clerical review.
4. Select a sample of record pairs from each batch. The acceptance sampling scheme will define the size of the sample required. The sample should be selected at random without replacement.

5. The record pairs in each sample are examined clerically and each is assigned a link or non-link status as described in section 2.2.
6. Compare the number of links in each sample to the two critical values defined by the acceptance sampling scheme.
7. If the number of links observed in the sample is less than the lower sample threshold all the record pairs in the batch are assigned as confirmed non-links (except for any record pairs that may have been identified as links in the sample).
8. If the number of links observed in the sample is greater than the upper sample threshold all the record pairs in the batch are assigned as confirmed links (except for any records pairs that may have been identified as non-links in the sample).
9. If the number of links observed in the sample is between or equal to the thresholds all the non-sampled record pairs in the batch are sent to clerical review.
10. Steps 5 to 9 are repeated for batches, until a determination can be made on every batch.

It may not be necessary to review each batch. Starting the review with a batch in the intermediate weight range and working up and down the weight range can reduce the number of batches that need to be reviewed. If a number of consecutive batches are sentenced as links then batches of a higher weight range are very likely to also be sentenced as links. Similarly, if a number of consecutive batches are sentenced as non-links then batches of a lower weight range are very likely also to be sentenced as non-links.

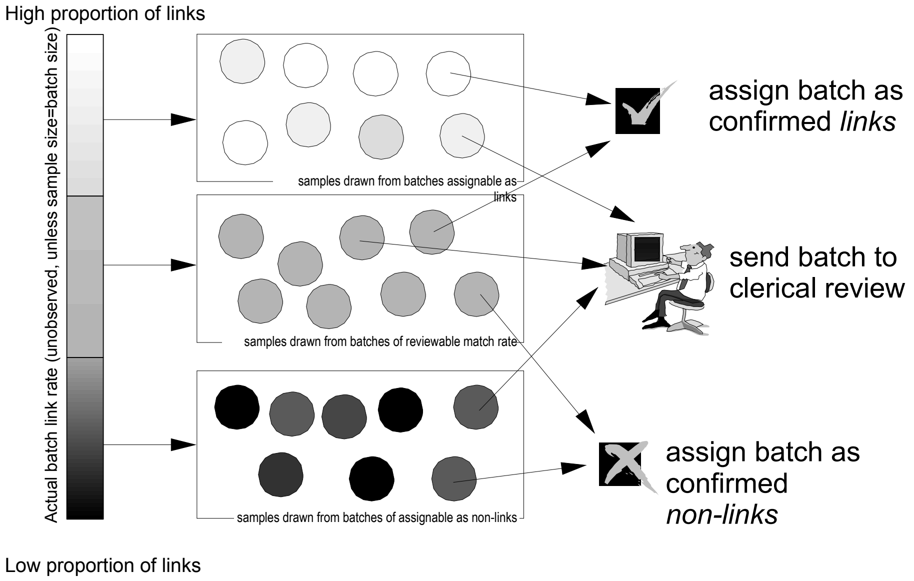
3.4 Possible errors in acceptance sampling clerical review

In any situation involving sampling, there is the possibility of making errors. Similarly, a number of errors are possible in acceptance sampling based clerical review. These errors are described below and also are shown diagrammatically in figure 3.1.

Suppose that a batch contains a high proportion of links, yet the sample drawn from the batch contains (by chance) a relatively low number of links. This batch may be assigned to clerical review, thus unnecessarily increasing the clerical review load. It is exceedingly unlikely that such a batch would be assigned as non-links.

Conversely, suppose that a batch contains a low number of links, but the drawn sample contains (by chance) a relatively high number of links. We may send this batch to clerical review, thus unnecessarily increasing the clerical review load. It is exceedingly unlikely that such a batch would be assigned as links.

3.1 Possible outcomes from the clerical of batches


High proportion of links
Actual batch link rate (unobserved, unless sample size=batch size)
samples drawn from batches assignable as links
samples drawn from batches of reviewable match rate
samples drawn from batches of assignable as non-links
assign batch as confirmed *links*
send batch to clerical review
assign batch as confirmed *non-links*
Low proportion of links

A sample drawn from a batch that should really be clerically reviewed may have a relatively higher or lower number of links and we may assign the batch as links or non-links, increasing the rate of linkage errors.

3.5 Designing an acceptance sampling process

The risk of errors such as those described in Section 3.4 may be mitigated by careful design of the acceptance sampling process. This requires the informed specification of the following process attributes:

1. a sample size – the number of record pairs to be selected from each batch;
2. a lower threshold – if the number of links observed in the sample is less than this threshold, the whole batch is assigned as non-links;
3. an upper threshold – if the number of links observed in the sample exceeds this threshold, the whole batch is assigned as links;
4. a clerical review region – if the number of links observed in the sample lies between the two threshold values (inclusive of the threshold values) then the batch is assigned to clerical review;
5. a mechanism for allocating record pairs to batches.

Tables for generating acceptance sampling schemes can be found in Australian or International standards, such as AS 1199.0-2003 or ISO 2859-10:2003 (Sampling Procedures for Inspection by Attributes).

Statistical software packages also have procedures to construct acceptance sampling schemes, for example PROC POWER in SAS.

An acceptance sampling scheme can be designed by specifying the following parameters:

1. the risk of assigning a batch containing a low proportion of matches to clerical review;
2. the proportion of matches that is considered to be a ‘low proportion’;
3. the risk of assigning a batch containing a high proportion of matches to clerical review;
4. the proportion of matches that is considered to be a ‘high proportion’;
5. the risk of assigning a batch as non-links that really should be assigned to clerical review;
6. the risk of assigning a batch as links that really should be assigned to clerical review;
7. the lower limit of the proportion of matches in a batch that should be assigned to clerical review;
8. the upper limit of the proportion of matches in a batch that should be assigned to clerical review; and
9. optionally, the batch size.

3.2 Example of a clerical review acceptance sampling scheme

Sample Size	300		
Tolerances	**Lower**		**Upper**
Minimum match rate for clerical inspection	0.3	Acceptable match rate for automatic links	0.9
	90		270
Risk of having to review a batch with a large proportion of non-matches	0.05	Risk of having to review a batch with a large proportion of matches	0.05
Critical Limit	0.3467	Critical Limit	0.8667
Sample Threshold	104	Sample Threshold	40
Clerical Review Region	104-260	links observed in sample	
Power	85%	Power	80%
Risk of assigning a batch with a reviewable proportion of non-matches as non-links	15.0%	Risk of assigning a batch with a reviewable proportion of matches as links	20.0%
Minimum Detectable Non-Compliance	37.3%	Minimum Detectable Non-Compliance	84.97%

Generally, controlling for some risks will require exceeding the constraints specified for other risks. It is recommend that visualisation tools are used to examine the properties of candidate acceptance sampling schemes. Table 3.2 shows an extract from a spreadsheet used to design and visualise an acceptance sampling scheme.

In the example in table 3.2 the following parameters have been specified:

- the risk of assigning a batch containing a low proportion of matches to clerical review (5%);
- the proportion of matches that is considered to be a 'low proportion' (30%);
- the risk of assigning a batch containing a high proportion of matches to clerical review (5%);
- the proportion of matches that is considered to be a 'high proportion' (90%);
- the risk of assigning a batch as non-links that really should be assigned to clerical review (15%); the complement of this risk is the probability of correct assignment to clerical review, this probability is denoted as power;
- the risk of assigning a batch as links that really should be assigned to clerical review (20%); the complement of this risk is the probability of correct assignment to clerical review, this probability is denoted as power;
- sample size (300).

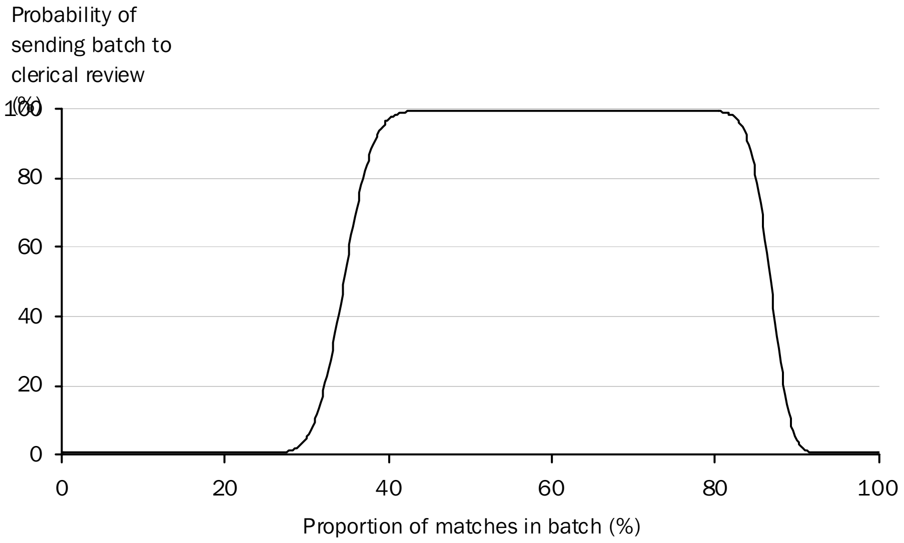
The spreadsheet generates:

- the lower limit of the proportion of actual matches in the whole batch that will be assigned to clerical review with a probability of at least the specified power (0.35);
- the upper limit of the proportion of actual matches in the whole batch that will be assigned to clerical review with a probability of at least the specified power (0.87);
- critical values of number observed links in a sample, used to assign batches (104 and 260).

The spreadsheet also generates a clerical review power curve. This curve plots the probability of sending a batch to clerical review as a function of the proportion of actual matches in the batch. This curve can be used in a similar manner as the Operating Characteristic Curve that is commonly used in statistical quality control. The Operating Characteristic curve plots the probability of accepting batches, which in the context of clerical review is the probability of assigning the batch as links or non-links. The clerical review power curves plots the probability of the complementary event of sending a batch to clerical review (i.e. 'rejecting' the batch).

This was used in preference to the probability of acceptance as a batch that should be accepted as links could in fact be accepted as non-links (and vice versa). Figure 3.3 shows the clerical review power curve generated by the sampling scheme specified in table 3.2.

3.3 Example clerical review power curve



3.6 Establishing cut-offs using sampling

Sampling offers an objective method of setting cut-offs. The sampling schemes described in Section 3.5 can be used to evaluate batches instead of automatically sentencing them. That is, the batch and sample sizes have been designed to assign a link status with confidence and can also be used to produce a reasonable estimates of links and non-links in each batch. These estimates can be used to set cut-offs. Effectively, this replicates the traditional decision rule, but sampling is used to set the cut-offs informatively and rigorously.

The record pairs are divided into batches with similar comparison weights, as before, and the proportions of links and non-links in each batch sample, determined by clerical review, are recorded. These proportions are multiplied by batch size to estimate the numbers of links and non-links in the whole batch. These estimates reflect the number of correct links (linked records presumed to be matches) and incorrect links (linked records presumed to be non-matches). Note that clerical review does not necessarily provide the true match status of a record pair, rather the most appropriate judgement given the data fields available on each file.

Starting at the maximum weight batch and working down, the estimated numbers of correct links and incorrect links as each successive batch are accumulated. These accumulated estimates can be used to make determinations on where to set cut-offs by optimising the number of incorrect links that would be accepted and the number of correct links that would be missed. The balance between missing correct links and incorporating incorrect links will depend upon the nature of the linkage. Bishop (2009) discusses the likely quality of linked datasets using a variety of cut-offs in a specific data linking application.

3.4 Sample clerical review result output

```
No record pairs: 35061   No Batches: 55   Sample Size: 65
 Batch#     Status       Min. weight    Max. weight        Size       Sample Size  Est. Defective
    0       N/A              57.63          58.63           457           65
    1       N/A              56.64          57.62           128           65
    2       N/A              55.64          56.61           116           65
    3       N/A              54.63          55.62            78           65
    4       N/A              53.63          54.59            74           65
    5       N/A              52.64          53.52          1821           65
    6       N/A              51.63          52.62           598           65
    7       N/A              50.63          51.62           333           65
    8       N/A              49.63          50.62           171           65
    9       N/A              48.64          49.62           201           65
   10       N/A              47.63          48.62           167           65
   11       N/A              46.65          47.62           299           65
   12       N/A              45.63          46.62           142           65
   13       N/A              44.65          45.62           610           65
   14       N/A              43.64          44.60           281           65
   15       N/A              42.67          43.61           173           65
   16       N/A              41.63          42.59           171           65
   17       N/A              40.63          41.62           152           65
   18       N/A              39.65          40.61           252           65
   19       N/A              38.64          39.62           108           65
   20       N/A              37.63          38.62           197           65
   21       N/A              36.63          37.61           158           65
   22       N/A              35.63          36.62           169           65
   23       N/A              34.63          35.62           287           65
   24       N/A              33.63          34.60           217           65
   25       N/A              32.65          33.62           177           65
   26       N/A              31.63          32.61           264           65
   27       N/A              30.63          31.62           152           65
   29       N/A              28.63          29.61           660           65
   30       N/A              27.63          28.62           253           65
   31       N/A              26.63          27.60           394           65
   32       N/A              25.63          26.62           204           65
   33       N/A              24.63          25.62           295           65
   34       N/A              23.63          24.61          7145           65
   35       N/A              22.63          23.62          2361           65
   36       N/A              21.63          22.62          2056           65
   37       N/A              20.63          21.62          1174           65
   38       confirmed        19.63          20.61           806           65              1.5%
   39       N/A              18.63          19.62          1723           65              1.5%
   40       N/A              17.63          18.62           664           65              3.1%
   41       N/A              16.63          17.63           405           65              3.1%
   42       N/A              15.63          16.61          1337           65              1.5%
   43       N/A              14.63          15.62           623           65              7.7%
   44       confirmed        13.63          14.62           895           65             13.8%
   45       confirmed        12.63          13.62           408           65             15.4%
   46       clerical         11.64          12.60           358           65             20.0%
   47       clerical         10.63          11.61           320           65             30.8%
   48       clerical          9.63          10.62           366           65             35.4%
   49       clerical          8.63           9.62           418           65             58.5%
   50       rejected          7.63           8.62           543           65             64.6%
   51       rejected          6.63           7.62           613           65             69.2%
   52       rejected          5.63           6.62           573           65             83.1%
   53       N/A               4.63           5.62           667           65
   54       N/A               3.63           4.62           849           65
```

Setting an upper and lower cut-off

An upper and lower cut-off can be set. All batches above the upper cut-off are accepted as links. All batches below the lower cut-off are accepted as non-links. Record pairs between the lower and upper cut-offs are submitted for clerical review. Setting a lower and upper cut-off is appropriate when there are fields available that are well suited for clerical review.

Example

In the example illustrated in table 3.4, record pairs were divided into batches of equal comparison weight ranges, where each weight range had width 1.0. A sample of 65 record pairs was randomly selected from each batch. Each of the record pairs was clerically examined to determine whether it was a link or non-link. The 'Est Defective' column indicates the estimated proportion of non-links in each batch, and is simply the proportion of non-links in the sample. The estimated number of non-links in each batch is compared to the lower and upper thresholds of the clerical review region to determine the status of each batch. The status for each batch has been displayed in the 'Status' column. As noted in section 3.3, it is not necessary to sample every batch. In this example, the results of batches 38 to 45 are used to assume that batches 0 to 37, which will all be linked, also have a tolerable proportion of non-matches. Similarly, batches 53 and 54 were assumed to have a low proportion of matches based on the results of batches 50 to 52. This established a clerical review range consisting of record pairs with record pair comparison weight of 8.63 to 12.60 (representing batches 46 to 49).

Setting a single cut-off

A single cut-off can also be set. All batches above this single cut-off are accepted as links. All batches below the cut-off are accepted as links. Setting a single cut-off is appropriate where the data fields are less suited for clerical review, for example linkages that do not use name and address.

Example

Table 3.5 shows an example of sampling clerical review output. In this example samples of size 65 were selected from batches of record pairs with comparison weights in equal weight ranges. The estimated proportion of incorrect links in each batch is equal to the observed proportion of incorrect links in the sample. This estimate is applied to the batch size to estimate the numbers of incorrect and correct links in each batch. These batch level estimates have been accumulated by decreasing record pair comparison weight. The results were also examined graphically by plotting the estimate of incorrect and correct link against record pair comparison weight, see figure 3.6.

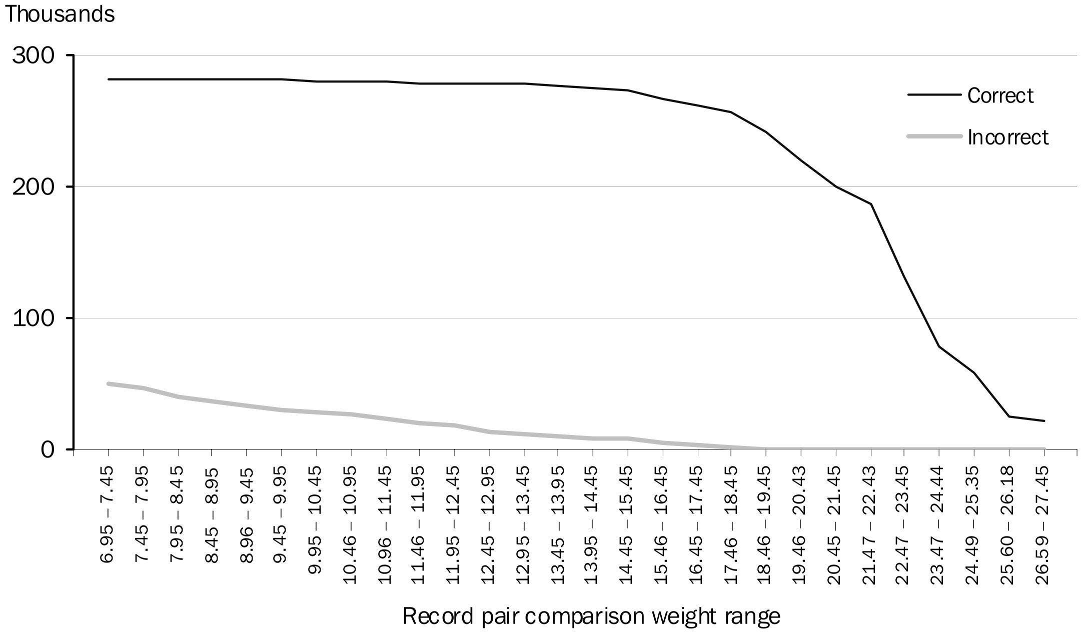
3.5 Sample clerical review result output used to select a single cut-off

Batch number	Weight range	Batch size	Sample size	Proportion of incorrect links	*Estimated number of links*			
					Per batch		*Cumulative*	
					Incorrect	*Correct*	*Incorrect*	*Correct*
0	26.59–27.45	21,140	65	0.0%	0	21,140	0	21,140
1	25.60–26.18	3,578	65	0.0%	0	3,578	0	24,718
2	24.49–25.35	33,772	65	0.0%	0	33,772	0	58,490
3	23.47–24.44	18,783	65	0.0%	0	18,783	0	77,273
4	22.47–23.45	53,444	65	0.0%	0	53,444	0	130,717
5	21.47–22.43	55,268	65	0.0%	0	55,268	0	185,985
6	20.45–21.45	14,098	65	1.5%	211	13,887	211	199,872
7	19.46–20.43	19,176	65	0.0%	0	19,176	211	219,048
8	18.46–19.45	23,167	65	0.0%	0	23,167	211	242,215
9	17.46–18.45	16,217	65	9.2%	1,492	14,725	1,703	256,940
10	16.45–17.45	4,928	65	13.8%	680	4,248	2,383	261,188
11	15.46–16.45	6,392	65	30.8%	1,969	4,423	4,352	265,611
12	14.45–15.45	10,797	65	32.3%	3,488	7,309	7,840	272,920
13	13.95–14.45	2,070	65	44.6%	923	1,147	8,763	274,067
14	13.45–13.95	3,559	65	33.8%	1,203	2,356	9,966	276,423
15	12.95–13.45	2,662	65	60.0%	1,597	1,065	11,563	277,488
16	12.45–12.95	2,586	65	89.2%	2,307	279	13,870	277,767
17	11.95–12.45	4,966	65	93.8%	4,658	308	18,528	278,075
18	11.46–11.95	2,384	65	86.2%	2,055	329	20,583	278,404
19	10.96–11.45	3,567	65	78.5%	2,800	767	23,383	279,171
20	10.46–10.95	2,905	65	93.8%	2,725	180	26,108	279,351
21	9.95–10.45	2,795	65	61.9%	1,730	1,065	27,838	280,416
22	9.45–9.95	3,099	65	61.9%	1,918	1,181	29,756	281,597
23	8.96–9.45	3,630	65	83.0%	3,013	617	32,769	282,214
24	8.45–8.95	3,961	65	97.5%	3,862	99	36,631	282,313
25	7.95–8.45	3,906	65	100.0%	3,906	0	40,537	282,313
26	7.45–7.95	5,160	65	100.0%	5,160	0	45,697	282,313
27	6.95–7.45	3,558	65	100.0%	3,558	0	49,255	282,313

In this example three cut-offs were selected using this output to produce a low, medium and high linking standard. For example, 12.95 was selected for the low standard cut-off, resulting in approximately 277,500 correct links and 11,500 incorrect links. In contrast, 21.47 was used for the high cut-off, resulting in approximately 186,000 correct links and no incorrect links.

If more accurate cut-offs are desired, further sampling using narrower batch ranges can be employed around the cut-offs.

3.6 Estimated cumulative record pairs correctly and incorrectly linked by cut-off weight range



4. IMPLEMENTATION IN FEBRL

The data linking software chosen for the Census Data Enhancement project was FEBRL (Christen, Churches and Hegland, 2004). FEBRL 0.3 was released under an open source licence and was modified significantly in the ABS. Version 0.3 implements the Felligi–Sunter classifier, allowing for each record pair to be classified as a *link*, *non-link* or *possible link*. While FEBRL 0.3 could identify possible links, i.e. those requiring clerical review, there was no provision within the software to undertake this review.

The modifications for by the ABS introduced a facility for clerical review, as well as the capacity to undertake clerical review on an acceptance sampling basis. The ABS also made modifications so that the clerical review load could be split up amongst a number of clerical review staff.

Before running either the basic clerical module or the sampling based clerical module, the existing module to allocate cut-offs is run. The user specifies a lower clerical cut-off below which the record pairs are assigned as non-links and an upper clerical cut-off above which record pairs are assigned as links. Record pairs between, or equal to, the lower and upper cut-off are flagged for clerical review. Clerical review will be run on a completely enumerated basis or sample basis depending on which module is run next.

The basic clerical review module displays the data fields for record pairs assigned as possible links from the two files side by side to allow the reviewer to make an assessment of link status. A simple command line interface displays the data fields on the screen and provides a command prompt that allows the reviewer to assign a link status. The command prompt also allows the reviewer to navigate through record pairs.

The sampling based clerical review module is an extension of the basic clerical review module. The user can specify the sampling scheme in a number of ways. The number of records pairs in each batch can be selected in a number of ways, by specifying one of:

- the number of batches of clerical review record pairs;
- the percentage of total clerical review pairs to include in each batch;
- the number of clerical record pairs per batch; or
- the record pair comparison weight width of batches.

The number of record pairs to sample in each batch can be selected either by specifying:

- a fixed sample size per batch, or
- a percentage of record pairs to select from each batch, in which case the user can optionally specify a minimum and maximum sample size.

For acceptance sampling based clerical review, critical values are also specified. The critical values for accepting batches as links or non-links are selected by specifying either the critical number of links observed in sample or a critical proportion of links observed in sample. The acceptance sampling schemes described in section 3.5 use a fixed sample size per batch and critical limits defined in determines of counts (rather than proportions) of links observed.

Additionally the user can specify a run number so that it is possible to review which run during a multi-run linkage project a record pair status was determined.

Once the sampling parameters have been selected a command line interface allows the reviewer to navigate through batches and review summary results. Once a batch has been navigated to, the user is presented with a similar command line interface as basic clerical review. The clerical review interface provides a running total of the number and proportion of links and non-links identified, as well as providing options for sentencing batches once one of the critical limits have been reached or to continue reviewing the batch. The former option is recommended for acceptance sampling based clerical review and the latter option for when sampling is used to establish cut-offs (or a single cut-off).

5. DISCUSSION

Issues faced by data linking practitioners include determining the appropriate amount of clerical review to perform and setting reliable cut-offs. Too much clerical review can be burdensome and time consuming. Too little clerical review can result in reduced linkage quality by missing links or including incorrect links. The sampling methods described in this paper provide a framework for establishing cut-offs in an informed manner to optimise the amount of clerical review.

Further reductions in clerical review are possible by using acceptance sampling methods. Missed links and incorrect links are inevitable in probabilistic linking. The acceptance sampling method can increase both missed links and incorrect links. However the extent of missed links and incorrect links can be controlled by selecting a sampling scheme appropriate for the linkage objectives.

Sampling methods have been used successfully during a number of studies during the Census Data Enhancement Project, for example Solon and Bishop (2009) and Wright (2010). The main benefit of sampling was found to be in establishing a profile of the estimated links and non-links by record pair comparison weight to establish effective cut-offs.

In the simulated formation of the Statistical Longitudinal Census Dataset, initial clerical review bounds for the first pass would have yielded 10,828 clerical review pairs. These initial bounds were based on the distribution of comparison weights. Sampling was used to establish revised cut-offs. This reduced the number of clerical review pairs to approximately 4,400.

Further enhancements can be made to the sampling procedures used. For example, during FEBRL implementation, a *double sampling* option was assessed. Double, multiple and sequential acceptance sampling schemes have the potential to produce even greater reductions in clerical review. However, it was found that such sampling schemes can compromise the capacity to generate a reliable profile of links and non-links by comparison weight. This strongly influenced the decision to discontinue further development of these options at the present time.

ACKNOWLEDGEMENT

The author wishes to acknowledge the outstanding effort and professional service provided by Sarah Healy in programming the implementation of the sampling and clerical review modules.

REFERENCES

Australian Bureau of Statistics (2005a) *Census Data Enhancement – Statement of Intention*, (last viewed on 9 May 2011) <http://www.abs.gov.au/websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/5812a287d6a2e78fca2571ee001a7a49!OpenDocument>

—— (2005b) *Enhancing the Population Census: Developing a Longitudinal View*, Discussion Paper, cat. no. 2060.0, ABS, Canberra.

—— (2006) *Census Data Enhancement Project: An Update*, Information Paper, cat. no. 2062.0, ABS, Canberra.

Australian Institute of Health and Welfare (2003) *Interface between Hospital and Residential Aged Care: Feasibility Study on Linking Hospital Morbidity and Residential Aged Care Data*, (last viewed on 9 May 2011) <http://www.aihw.gov.au/publication-detail/?id=6442467495>

Bishop, G. (2009) "Assessing the Likely Quality of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Centre for Health Record Linkage (CHeReL) (2008) *Quality Assurance in Record Linkage*, (last viewed on 9 May 2011) <http://www.cherel.org.au/CHeReLQualityAssuranceJuly2008.pdf >

Christen, P., Churches, T. and Hegland, M. (2004) "Febrl - A Parallel Open Source Data Linkage System", *Proceedings of the 8th Pacific-Asia Conference*, PAKDD 2004, Sydney.

Christen, P. and Churches, T. (2005) *FEBRL 0.3 Documentation*, (last viewed on 9 May 2011) <http://cs.anu.edu.au/~Peter.Christen/FEBRL/FEBRL-0.3/FEBRLdoc-0.3/>

Christen, P. (2008) "FEBRL – A Freely Available Record Linkage System with a Graphical User Interface", in James R. Warren, Ping Yu and John Yearwood (ed.), *Conferences in Research and Practice in Information Technology, Volume 80*.

Conn, L. and Bishop, G. (2006) "Exploring Methods for Creating a Longitudinal Census Data Set", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Gill, L. (2001) "Methods for Automatic Record Matching and Linkage and their Use in National Statistics", *National Statistics Methodological Series*, No. 25, Office for National Statistics, London.

Juran, J.M. and Godfrey, A.B. (1999) *Juran's Quality Handbook*, Fifth edition, McGraw Hill, New York.

Karmel, R. (2004) *Linking Hospital Morbidity and Residential Aged Care Data: Examining Matching due to Chance*, cat. no. AGE 40, AIHW, Canberra, (last viewed on 9 May 2011) <http://www.aihw.gov.au/publications/index.cfm/title/10065>

Montgomery, D.C. (2005) *Introduction to Statistical Quality Control*, Fifth edition, Wiley, New York.

Porter, E.H. and Winkler, W.E. (1997) "Approximate String Comparison and its Effect on an Advanced Record Linkage System", *Research Report RR97/02*, U.S. Bureau of the Census.

Solon, R. and Bishop, G. (2009) "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Wright, J. (2010) "Linking Census Records to Death Registrations", *Methodology Research Papers*, cat. no. 1351.0.55.030, Australian Bureau of Statistics, Canberra.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

© Commonwealth of Australia 2011
Produced by the Australian Bureau of Statistics